

AD-A113 464

AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TX
ENLISTMENT SCREENING TEST FORMS 81A AND 81B: DEVELOPMENT AND CA--ETC(U)
MAR 82 M J REE
AFHRL-TR-81-54

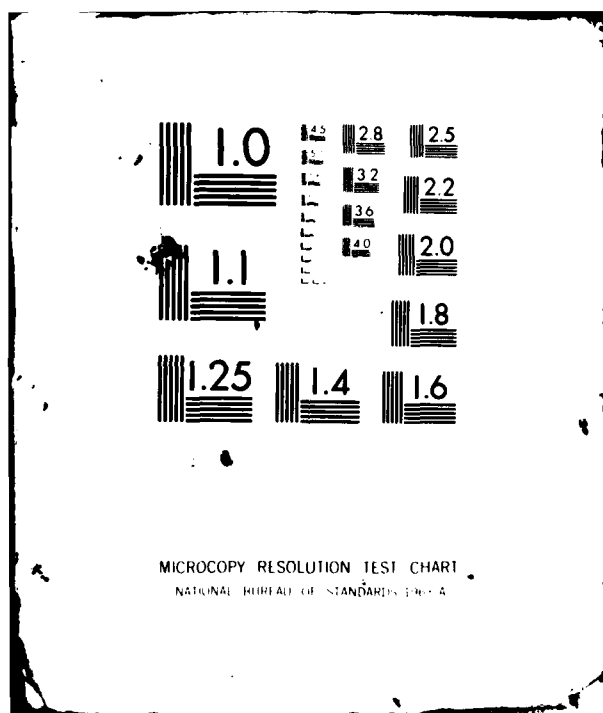
F/G 5/9

UNCLASSIFIED

NL

1 1 1
AD 140.2

END
DATE
FILMED
5-82
DTIC



AIR FORCE



HUMAN RESOURCES

AD A113464

DTIC FILE COPY

**ENLISTMENT SCREENING TEST
FORMS 81a AND 81b:
DEVELOPMENT AND CALIBRATION**

By

**John J. Mathews
Malcolm James Ree**

**MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235**

March 1982

Final Report

Approved for public release; distribution unlimited.

**DTIC
ELECTE
APR 15 1982**

S

E

LABORATORY

**AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235**

82 04 15 032

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

NANCY GUINN, Technical Director
Manpower and Personnel Division

RONALD W. TERRY, Colonel, USAF
Commander

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

Item 20 (Continued):

equated) to the Armed Forces Qualification Test (AFQT) through the method of equivalent percentiles. The ESTs appear to be highly reliable instruments, discriminating well throughout a range which includes major service selection cutoff points. The two EST forms appear parallel based on highly similar distributions of item difficulty and criterion correlation values. EST scores predict AFQT percentiles quite well ($r = .83$). In addition, EST content is similar to that of AFQT.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

PREFACE

This study was completed under Task 771918, Selection and Classification Technologies, which is part of a larger effort in Force Acquisition and Distribution. It was subsumed under work unit number 77191804, "Maintenance and Improvement of Enlistment Selection and Classification Tests," and executed as part of the Air Force Human Resources Laboratory (AFHRL) responsibility as lead laboratory under the executive agent (USAF) for Armed Services Vocational Aptitude Battery research and development.

The authors wish to express their appreciation to Richard Kotula, Jim Friemann, and Roy Chollman of AFHRL for their assistance during this project.

Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	



TABLE OF CONTENTS

	Page
I. Introduction	5
II. Method	5
Construction of Experimental Booklets	5
Administration of Booklets	5
Data Editing	5
Samples	6
Data Analysis	6
III. Results and Discussion	7
Comparison of EST Forms	7
Calibration of EST 81a	8
Relationship of EST to General Composites	10
IV. Summary, Conclusions, and Recommendation	10
References	10
Appendix A: Statistical Tables	11

LIST OF TABLES

Table	Page
1 Content of AFQT and EST Forms 81a-81b	7
2 Distributions of Item Difficulties for EST Forms 81a and 81b	7
3 Distributions of EST Item Validities with AFQT	7
4 Means of IRT Item Parameters for EST Form 81a	8
5 Descriptive Statistics and Correlations for EST Form 81a in an AFQT-Stratified Sample (N = 486)	8
6 Equipercentile Calibration of EST to AFQT (N = 869)	9
A1 Distribution of Demographic Variables by Subsample	11
A2 Distribution of EST 81a Scores by AFQT Category	12

ENLISTMENT SCREENING TEST FORMS 81a AND 81b: DEVELOPMENT AND CALIBRATION

I. INTRODUCTION

The Enlistment Screening Test (EST) was developed at the request of the military recruiting commands to reduce enlistment processing costs for transportation and boarding associated with testing of applicants. By administering this test at local recruiting stations, those applicants who would most likely meet service mental qualification standards could be identified and would be sent to centralized testing stations.

Previous EST Forms 5 and 6 (Jensen & Valentine, 1976) were designed to predict qualification on the Armed Forces Qualification Test (AFQT) portion of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 5, 6, and 7. These ESTs became obsolete with the implementation of the new ASVAB forms which do not have Space Perception items in the AFQT portion of ASVAB as did prior forms.

The objective of this study was to develop and provide norms for two parallel EST forms for use by military recruiters in predicting applicant success on ASVAB Forms 8, 9, and 10 selection composites. The new ESTs were to be designed in accordance with specifications which would make them appropriate for use by all armed services.

II. METHOD

Construction of Experimental Booklets

Two parallel forms of an EST, each of which would require no more than 45 minutes to administer, were to be developed (the operational AFQT to be predicted takes about 75 minutes). Three experimental booklets containing 60 items each and requiring 1 hour of administration time were assembled. Three booklets, rather than two, were constructed so that testing time would be shortened and fatigue effects would be lessened. Eight items were common to all three booklets as a check on the comparability of samples administered the different booklets.

All AFQT item types except Numerical Operations (NO) were to be used in the ESTs. Because of administration difficulties, it was decided to exclude the speeded NO item types from the EST. The total pool of items consisted of 74 Word Knowledge (WK), 60 Arithmetic Reasoning (AR), and 30 Paragraph Comprehension (PC) items, all of which were unique items. These items were selected based on available data that indicated they would discriminate best in the range of the 5th to 55th percentile for samples stratified on AFQT scores. The goal for final EST item selection was to have about 40% of the items peaked in difficulty around the 15th percentile level, 40% peaked in difficulty around the 30th percentile, and the other 20% peaked in difficulty around the 45th percentile level. This distribution would maximize measurement reliability at ability levels where most selection decisions are made.

The three preliminary booklets were developed to allow for item analysis on AFQT-stratified samples preparatory for final item selection for two EST forms. Items from the three experimental booklets were to be selected to form two parallel operational EST forms, each containing between 40 and 50 items.

Administration of Booklets

One of three experimental test booklets were individually administered to service applicants at recruiting stations during March and April 1981. A total of 1,853 subjects tested at about 300 stations had sufficient data forwarded in time to be included in the various analyses. Service composition of the applicants was Army, 33%; Navy, 29%; Air Force, 18%; and Marine Corps, 20%.

Data Editing

The answer sheets received were manually screened for completeness. The following steps were taken to insure that only valid data were included in analyses:

1. Correct designation of booklet form on the answer sheet was checked by scoring the test with the answer key for each of the three booklets. When an appreciably higher score was obtained based on the key for a form other than that designated on the answer sheet, the booklet designation was changed. When the highest of the three scores was below eight (out of 60 items), indicating that the examinee was not trying on the test, the case was deleted from the study.

2. Cases with no AFQT scores were deleted from all analyses except preliminary item analyses.

3. Cases with a standard error of estimate greater than +3.5 points from the predicted AFQT based on EST scores were deleted from the study. Less than one case per 1,000 would be expected by chance to lie outside of this range.

Fewer than 2% of the cases were eliminated based on these editing procedures.

Samples

The following subsamples were formed for specific analyses:

Sample 1. Subjects given experimental booklet AX (N = 527) and for whom answer sheets were received by 24 April 1981.

Sample 2. Subjects given experimental booklet BY (N = 486) and for whom answer sheets were received by 24 April 1981.

Sample 3. Subjects given experimental booklet CZ (N = 457) and for whom answer sheets were received by 24 April 1981.

Sample 4. Subjects given experimental booklet BY (N = 869) and for whom answer sheets were received by 14 May 1981. Sample 2 is a subset of Sample 4.

Data Analysis

Item statistics were generated to aid item selection for operational forms. In order to print final tests as soon as possible, item analysis was limited to cases available by 24 April 1981 (Samples 1 to 3). Descriptive statistics including correlations and bivariate frequency distributions of EST with ASVAB selection composites were computed. The EST was calibrated to AFQT through equipercentile equating.

Samples 1 through 3 were rectilinearly stratified on AFQT percentile by random duplication of subjects so that an equal portion (10%) of the sample was represented in each AFQT decile. This procedure equated the samples on ability. Classical item analysis, including item difficulty and discrimination indices with AFQT as an external criterion, was then accomplished for each experimental form.

Items were selected to yield two parallel EST forms. All the items for one operational form were chosen from the booklet (BY) which contained the greatest number of items in the appropriate difficulty and discrimination ranges. This procedure subsequently allowed direct generation of EST total score frequency distribution and bivariate statistics involving EST and AFQT, since a portion (sample 4) of the subjects was given all items which would be in one operational EST form.

Considerations in selecting items for the two final forms included:

1. Significant positive correlation with AFQT.
2. Maximum discrimination in the desired ability range (5th to 55th percentile)
3. Content similar to AFQT (proportion of Verbal and Arithmetic Reasoning items).
4. Necessity of having the two forms parallel.

III. RESULTS AND DISCUSSION

Comparison of EST Forms

Two EST forms of 48 items each which met the desired specifications were constructed. A comparison of the content of these forms, designated EST 81a and 81b, and AFQT is given in Table 1. After deleting the Numerical Operations items, AR items equally comprise 37.5% of both AFQT and EST content. There are relatively more WK and fewer PC items in EST than in AFQT. However, these two verbal item types are highly intercorrelated (Ree, Mullins, Mathews, & Massey, 1982). Because WK items require less time to complete, an abundance of these items saves testing time.

Table 1. Content of AFQT and EST Forms 81a and 81b

Item Types	Number of Items	
	AFQT	EST
Arithmetic Reasoning (AR)	30	18
Word Knowledge (WK)	35	23
Paragraph Comprehension (PC)	15	7
(Verbal = WK + PC)	(50)	(30)
Numerical Operations (NO)	50 ^a	—
Total Number of items	130	48

^aSpeeded test with 50 items. No raw scores are weighted .5 in the AFQT composite.

Difficulty (*p*) levels (proportion answering items correctly) of the two forms based on stratified samples are indicated in Table 2. Item *p*'s range from .57 to .89. Mean *p*'s for the ESTs are virtually identical (.744 and .742), and the distributions are similar. Biserial correlations (validity estimates) of items with AFQT percentiles ranged from .29 to .63, with about 75% between .45 and .59 (see Table 3). Again, the forms appear quite comparable, with similar means (.476 and .496) and distributions. An internal consistency reliability (Kuder-Richardson Formula 20, Lord & Novick, 1968) of .93 was obtained for EST 81a. Since no subjects took all items in EST 81b, test statistics, including reliability, could not be precisely computed for this form.

Table 2. Distributions of Item Difficulties for EST Forms 81a and 81b

Difficulty (<i>p</i>)	81a (Booklet BY)		81b (Booklets AX & CZ)	
	N	%	N	%
.80 - .89	12	25	12	25
.70 - .79	23	48	19	40
.60 - .69	12	25	16	33
.50 - .59	1	2	1	2
Mean <i>p</i>	.744		.742	

Table 3. Distributions of EST Item Validities with AFQT

Validity (r_{bis})	81a (Booklet BY)		81b (Booklets AX & CZ)	
	N	%	N	%
.60 - .99	2	4	3	6
.45 - .59	34	71	34	71
.30 - .44	11	23	11	23
.00 - .29	1	2	0	0
Mean $r_{biseria}$ ^a	.476		.496	

^aBased on *r* to *Z* transformations.

Item Response Theory (IRT) item analytic indices (Lord & Novick, 1968) were also computed for EST 81a based on the Birnbaum (1968) three-parameter logistic model as implemented in OGIVIA (see Ree, 1979). The three indexes are a (item discrimination), b (item difficulty in unit normal metric), and c (probability of guessing) (see Ree, 1979, for a detailed description of these item parameters). Two types of analyses were completed. The first was based on EST score as an internal criterion, and the second was based on AFQT percentiles transformed into Z scores for the sample and used as an external criterion. Table 4 presents the mean a , b , and c values for these analyses on EST 81a.

Table 4. Means of IRT Item Parameters for EST Form 81a

Analysis	\bar{a}	\bar{b}	\bar{c}
Internal Criterion	1.33	-.62	.20
AFQT Criterion	1.10	-.62	.32

As would be expected (see Buchmeier & Weiss, 1981, pp. 13-18), \bar{a} (the mean of a) is somewhat higher in the internal compared to external (AFQT) criterion analysis, although both \bar{a} values were relatively high (1.3 and 1.1, respectively). Both \bar{b} values were -.62, corresponding to a mean percentile of 27. The mean of the desired (targeted) distribution of b would be about -.67. The \bar{c} was .20 based on the internal analysis. The classical probability of guessing for a four-choice item is .25. The \bar{c} based on AFQT ability estimates was somewhat higher, .32, but it may have been poorly estimated due to the low item difficulty and small sample size (Ree & Jensen, 1980).

Table 5 gives the mean and standard deviation (SD) of EST 81a scales based on sample 2 stratified on AFQT. Summary statistics for EST 81b should be quite comparable, since the available data indicate it is indeed parallel to EST 81a.

Table 5. Descriptive Statistics and Correlations for EST Form 81a in an AFQT-Stratified Sample (N = 486)

Scale	Intercorrelations					
	Mean	SD	VE	AR	EST	AFQT
Verbal Ability (VE)	22.7	7.4	1.00	.69	.95	.77
Arithmetic Reasoning (AR)	13.0	4.7		1.00	.88	.75
EST Total	35.7	11.1			1.00	.83
AFQT Percentile	49.5	28.4				1.00

Intercorrelations are also shown in Table 5. The Pearson product-moment correlation (r) of EST 81a with the AFQT percentile is .83. This indicates that about 69% of the variance in AFQT scores can be predicted from EST 81a scores. The two EST subscales (Quantitative and Verbal) correlate with EST total to about the same degree as the like-named ASVAB composites correlate with AFQT. The r values of the Verbal (VE) and Arithmetic Reasoning (AR) scales with EST total are .95 and .88, respectively. The corresponding r values of ASVAB Forms 8, 9, and 10 VE and AR with ASVAB Forms 8, 9, and 10 AFQT are .93 and .89, respectively (Ree et al., 1981).

Calibration of EST 81a

An equipercentile calibration was accomplished to convert EST scores to equivalent AFQT percentiles. Table 6 gives the cumulative percent (Column 2) of sample 4 (N = 869) below each raw score on EST 81a and its percentile equivalent (Column 3) when calibrated against AFQT. Table 6 also shows cumulative percent (Column 4) falling below the standard percentile on AFQT.

Table 6. Equipercntile Calibration of EST to AFQT (N = 869)

(1) Raw EST Score	(2) Cumulative % Below EST Score	(3) AFQT Percentile	(4) Cumulative % Below AFQT Score
1-11	1.4	4	0.9
12	1.7	5	1.5
13	2.3	6	2.6
14	2.8	6	2.6
15	3.2	8	3.3
16	3.6	8	3.3
17	4.6	10	4.5
18	5.4	12	5.5
19	6.2	13	6.2
20	6.9	14	7.4
21	7.4	14	7.4
22	8.5	15	8.6
23	9.8	15	8.6
24	10.6	16	10.9
25	11.5	18	12.1
26	12.7	19	12.8
27	14.4	20	14.8
28	16.0	22	16.6
29	17.5	23	17.7
30	19.9	25	20.6
31	22.0	28	22.9
32	24.2	30	25.5
33	27.4	33	28.8
34	30.6	36	32.5
35	34.3	40	36.0
36	38.0	42	38.4
37	41.1	44	41.4
38	44.9	48	45.3
39	47.6	49	47.8
40	50.5	50	50.5
41	56.5	54	56.8
42	60.4	59	62.6
43	64.8	63	66.9
44	70.0	68	72.0
45	75.5	74	77.6
46	82.9	80	85.4
47	90.0	87	92.8
48	96.0	95	98.0

The EST score distribution is negatively skewed, indicating potential for good discrimination at lower ability levels. This was one of the test construction goals. The median EST score is 40 (out of 48 points), and this is equivalent to an AFQT percentile score of 50.

The percentages of subjects within various AFQT score categories are shown for EST scores (shown in intervals of 2 score points to increase Ns) in the appendix (Table A2). The AFQT categories represent various service cutoff score boundaries. The probability of obtaining an AFQT score above a category given various EST scores can be estimated by summing the percentages to the right of the category. For example, an applicant with an EST score of 21 or 22 would have about a 14% chance of obtaining an AFQT percentile of 31 or above.

Relationship of EST to General Composites

The correlation between the EST and the General-Technical (General for Air Force) composite for subjects given booklet BY in March 1981 ($N = 270$) was .86. This strong relationship was expected due to similarity in content. Since percentile equivalents for this composite are based on the same reference test used in norming AFQT and since it correlates highly with AFQT ($r \approx .97$), the equipercentile calibration (Table 6) would also apply to General (General-Technical) percentiles.

IV. SUMMARY, CONCLUSIONS, AND RECOMMENDATION

Two parallel forms of an EST have been developed, and one (EST 81a) has been calibrated to the AFQT selection composite from ASVAB Forms 8, 9, and 10. These tests appear to adequately meet administrative and psychometric specifications. All items correlate positively with total test score and AFQT scores and are in an appropriate range of difficulty (from average to very easy) for use in prescreening service applicants.

The EST appears to be a highly reliable instrument (internal consistency coefficient for EST 81a = .93). The test items appear to discriminate well throughout a range which includes major service selection cutoff points (AFQT percentiles 15 to 45). The two EST forms appear parallel based on highly similar distributions of item difficulty and criterion correlation values. EST scores predict AFQT percentiles quite well ($r = .83$). In addition, EST content is similar to that of AFQT.

Interpretation of EST scores is provided by an equipercentile calibration to AFQT (Table 6). This calibration also applies to prediction of General-Technical percentile scores.

It is recommended that EST 81a and 81b replace earlier EST forms for any prescreening of AFQT or General composites based on ASVAB Forms 8, 9, and 10.

REFERENCES

- Birnbaum, A. Some latent-trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.
- Buchmeier, J.L., & Weiss, D.J. *Final report: Task orders 1, 2, 3, and 5*. Contract N00123-79-1273. Minneapolis: University of Minnesota, Department of Psychology, August 1981.
- Jensen, H.E., & Valentine, L.D., Jr. *Development of the Enlistment Screening Test-EST Forms 5 and 6*. AFHRL-TR-76-42, AD-A033 303. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, May 1976.
- Lord, F., & Novick, M. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.
- Ree, M. J. Estimating item characteristic curves. *Applied Psychological Measurement*, 1979, 3, 371-385.
- Ree, M.J., & Jensen, H.E. Effects of sample size on linear equating of item characteristic curve parameters. In O. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometrics Program, 1980.
- Ree, M.J., Mullins, C.J., Mathews, J.J., & Massey, R.H. *Armed Services Vocational Aptitude Battery: Item and factor analyses of forms 8, 9, and 10*. AFHRL-TR-81-55. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory, March 1982.

APPENDIX A. STATISTICAL TABLES

Table A1. Distribution of Demographic Variables by Subsample

Form of EST	EST Form					
	AX (N = 527)		BY (N = 486)		CZ (N = 457)	
	N	%	N	%	N	%
Service						
Air Force	97	18	78	16	87	19
Army	155	30	121	25	122	27
Marine Corps	101	19	104	21	96	21
Navy	138	26	162	33	126	28
Unspecified	36	7	21	4	26	6
Education						
High School Grad ^a	318	60	316	65	280	61
GED	39	7	47	10	41	9
Non-HS Graduate	125	24	91	19	92	20
Unspecified	45	8	32	7	44	10
Area (Zip Code)						
0 (New England)	19	4	21	4	9	2
1 (Northeast)	64	12	63	13	70	15
2 (Atlantic)	81	15	62	13	60	13
3 (Southeast)	71	13	65	13	66	14
4 (Midwest)	63	12	62	13	53	12
5 (North Central)	10	2	5	1	12	3
6 (Midlands)	58	11	48	10	41	9
7 (South Central)	77	15	70	14	64	14
8 (Rocky Mountain)	16	3	15	3	14	3
9 (Pacific)	56	11	50	10	45	10
Unspecified	12	2	25	5	23	5

^aIncludes subjects currently in the 12th grade when tested.

Table A2. Distribution of EST 81a Scores by AFQT Category

EST Score	AFQT Category									
	1-15		16-20		21-30		31-49		50-64	
	N	%	N	%	N	%	N	%	N	%
1-12	18	90.0	2	10.0	—	—	—	—	—	—
13-14	7	87.5	—	—	—	—	1	12.5	—	—
15-16	6	50.0	3	25.0	3	25.0	—	—	—	—
17-18	9	64.3	3	21.4	2	14.3	—	—	—	—
19-20	5	50.0	4	40.0	1	10.0	—	—	—	—
21-22	5	23.8	7	33.3	6	28.6	2	9.5	1	4.8
23-24	4	26.7	5	33.3	3	20.0	3	20.0	—	—
25-26	4	16.0	6	24.0	5	20.0	6	24.0	3	12.0
27-28	3	11.1	6	22.2	6	22.2	9	33.3	3	11.0
29-30	5	12.8	6	15.4	12	30.8	12	30.8	2	5.1
31-32	2	4.3	4	8.5	11	23.4	22	46.8	6	12.8
33-34	5	8.3	4	6.7	18	30.0	23	38.3	7	11.7
35-36	1	1.7	2	3.4	8	13.6	27	45.8	18	30.5
37-38	1	1.8	1	1.8	11	19.3	22	38.6	18	31.6
39-40	—	—	1	1.3	5	6.5	31	40.3	23	29.9
41-42	—	—	—	—	—	—	22	30.6	24	33.3
43-44	—	—	—	—	2	2.2	10	10.8	34	36.6
45-46	—	—	—	—	—	—	3	2.4	21	16.7
47-48	—	—	—	—	—	—	—	—	6	6.9
N	75		54		93		193		166	
									288	
										869

